Evaluation von Persistent-Identifier-Systemen für Forschungsdaten an der Humboldt-Universität zu Berlin

Stand: Januar 2013

Autoren: Elena Šimukovič[1], Benjamin Thomack[2], Paul Vierkant[3], Dennis Zielke[4]

Preface

Dieses Dokument gibt den Stand vom Januar 2013 wieder. Bei der Evaluation von Persistent-Identifier-Systemen (PIS) für Forschungsdaten an der Humboldt-Universität zu Berlin wurden die Aspekte Kosten, Aufwand, Qualitätssicherung, Verbreitung und Mehrwerte untersucht. Bestehende PIS sind DOI, handle, URN, ARK, PURL, und OpenURL.
URN sind lediglich auf Publikationen ausgelegt[5]; ARK sind nicht weit genug verbreitet; PURL und OpenURL finden ebenfalls im Bereich der eindeutigen Identifizierung von Forschungsdaten keine verbreitete Anwendung. Des Weiteren ist der in der UB eingesetzte OpenURL-Link-Resolver “SFX” eine proprietäre Lösung von ExLibris und ganz allgemein scheint der Einsatz eines kommerziellen proprietären Produkts für den Zweck der persistenten Identifizierung im

[1] http://orcid.org/0000-0003-1363-243X
[2] Humboldt-Universität zu Berlin
[3] http://orcid.org/0000-0003-4448-3844
[4] http://orcid.org/0000-0002-5764-8874 ODER Humboldt-Universität zu Berlin
[5] Eine Ausnahme stellt WDCC/DKRZ dar, das DOI und URN gleichzeitig einsetzt. Auch DANS/EASY nutzt URNs für Forschungsdaten.

wissenschaftlichen Bereich eher ungeeignet, da durch die Bindung an einen einzelnen Hersteller immer ein problematisches Abhängigkeitsverhältnis entsteht.
PURL wird keiner weiteren Betrachtung unterzogen, da keine Weiterentwicklung vorgesehen ist (selbst auf deren eigener Webseite sind viele Verlinkungen zu wichtigen Dokumentationsteilen nicht mehr aktuell oder vorhanden). Außerdem ergab ein Test der DOI-Foundation, dass nur 57% der PI resolvingfähig sind[6]. Daher werden im Folgenden nur DOI und handle verglichen, da beide PIS für Forschungsdaten ausgelegt sind.

DOI (Digital Object Identifier)

Allgemeine Informationen

Das DOI-System mit bisher über 55 Millionen vergebenen Identifiern für wissenschaftliche Publikationen - darunter etwa 1,4 Mio. für Forschungsdaten[7] - ist ein global weit verbreitetes PIS, dass von Bibliotheken, Rechenzentren und Verlagen gleichermaßen unterstützt wird.
Digital Object Identifier (DOI) für Forschungsdaten werden über das Konsortium DataCite vergeben. In Deutschland erfolgt die DOI-Vergabe für Institutionen über nach Fächern aufgeteilte Service-Agenturen, sobald man als Institution zu einem “Datenzentrum” wird:

- GESIS - Leibniz-Institut für Sozialwissenschaften,
- ZBW - Deutsche Zentralbibliothek für Wirtschaftswissenschaften,
- ZB MED - Deutsche Zentralbibliothek für Medizin,
- TIB - Technische Informationsbibliothek.[8]

Mit dem zwischen dem Datenzentrum und den Service-Agenturen geschlossenen Vertrag entstehen dem Datenzentrum derzeit keine Kosten. Das Datenzentrum erhält einen Präfix (z.B. 10.1392) und kann unbegrenzt DOI erstellen, indem es an dem Präfix unterschiedliche Suffixe anhängt. Für akademische Einrichtungen mit Hauptgeschäftssitz in Deutschland wird die DOI-Registrierung von der TIB kostenfrei vorgenommen[9].
Gleichzeitig verpflichtet sich das Datenzentrum zur Einhaltung des DataCite Metadatenschemas, dass folgende fünf Pflichtfelder voraussetzt:
Identifier (automatisch erstellte DOI), **Creator** (Name, Vorname), **Title**, **Publisher** (Name des Datenzentrums), **PublicationYear** (Jahr in dem der Datensatz vom Datenzentrum veröffentlicht wurde).

Kosten & Aufwand

Diese Verpflichtung zur formellen Erschließung der Forschungsdaten hat zur Folge, dass das Datenzentrum die Metadaten durch die WissenschaftlerInnen selbst oder durch Personal (z.B. Data Curator) erstellt lassen muss, wodurch wiederum personeller Aufwand und somit Folgekosten entstehen können. Ebenfalls können personelle Kosten für die technische

[6] http://nestor.sub.uni-goettingen.de/handbuch/artikel/nestor_handbuch_artikel_19.pdf
[7] Stand: 08.01.2013, nach http://stats.datacite.org/
[8] Bei multidisziplinären Datenzentren ist die TIB alleiniger Ansprechpartner.
[9] http://www.tib-hannover.de/fileadmin/doi/TIB-DOI-Kosten.pdf

Entwicklung der Schnittstelle zu DataCite entstehen.

Qualitätssicherung

Die “technische” Qualität der Metadaten wird durch die Übereinstimmung mit dem DataCite-Metadatenschema bzw. mit den bereits oben beschriebenen Pflichtfeldern gesichert. Für Speicherung/Hosting der Forschungsdaten und Aktualisierung der URL ist das Datenzentrum verantwortlich. Die Verantwortung für “inhaltliche” Qualität der Forschungsdaten bleibt dem Wissenschaftler überlassen.

Mehrwert

Der Mehrwert der Verpflichtung zur formellen Qualitätssicherung besteht bereits heute in unterschiedlichen Services, wie z.B. die Metadatensuche, Nutzungsstatistiken und die direkte Verlinkung auf die Forschungsdaten (LOD), die erst durch die zentrale Registrierung der DOI bei DataCite möglich gemacht werden.

Verbreitung

DOI werden z.B. von der LMU München (Software EPrints) oder Datadryad (DSpace) eingesetzt. Zu beiden Institutionen besteht bereits ein Austausch zu DOI.

handle

Allgemeine Informationen

Das DOI-System basiert technisch gesehen auf dem handle-System. Es gibt weltweit über 200.000 Präfixe, mit denen ähnlich wie bei DOI durch Suffixe Persistent Identifier erstellt werden können. Bis 2011 waren einzelne DOI kostenpflichtig, damals bot sich handle als kostengünstige Alternative an, sofern man eine große Anzahl an PI benötigte, z.B. zur Referenzierung von Millionen einzelner Datenpunkte in den Geowissenschaften.

Kosten & Aufwand

Um handles zu vergeben, registriert sich eine Institution bei handle.net. Mit dieser Registrierung entstehen jährliche Kosten in Höhe von derzeit 50$ wofür die Institution im Gegenzug einen Präfix erhält, mit dem sie durch Suffixe Persistent Identifier erstellen kann.
Technische Systeme wie DSpace und eSciDoc unterstützen out-of-the-box die Vergabe von handles. Für den Betrieb ist ein eigener lokaler handle-Server notwendig, wobei die Dauer der Verfügbarkeit dauerhaft sichergestellt werden muss. Hieraus entsteht ein erhöhter Personalaufwand, der sich aus Installation, Administration, Aufwand der periodischen Pflege der Identifier und Aufwand für die Anpassungen der jeweils angestrebten Repositorylösungen (z.B. LAUDATIO und Bilddatenbank) ergibt.

Qualitätssicherung

Es gibt bei handle.net weder ein auf Forschungsdaten zugeschnittenes Metadatenschema noch Pflichtfelder. Deshalb werden im Gegensatz zu DataCite die den handles zugehörigen

Metadaten auch nicht zu handle.net zurückgeschickt. Ebenfalls entstehen der einzelnen Institution aus der Registrierung bei handle.net keine vertraglichen Verpflichtungen zur formellen Qualitätssicherung. Es steht somit der Institution frei, ob und in welchem Maße Metadaten zu den Forschungsdaten angegeben werden müssen[10], wodurch ebenfalls die Kosten für die Metadatenerstellung reguliert werden können.

Mehrwert

Die mangelnde formelle Qualitätssicherung zusammen mit dem Fehlen eines zentralen Metadatenpools lassen jedoch keine Mehrwertdienste wie bei DataCite zu. handles dienen daher lediglich dem persistenten und eindeutigen Referenzieren bzw. globalen Resolving.

Verbreitung

Es gibt keine genaue Angabe, wieviele handles weltweit für Publikationen und Forschungsdaten vergeben sind.

Vorläufige Empfehlung

Technisch unterscheiden sich DOI und handle nicht. Hinsichtlich der Nutzung ist anzunehmen, dass DOI das weitverbreitetste und somit bei Wissenschaftlern bekannteste PIS für digitale Resourcen (Publikationen und Forschungsdaten) ist.[11] Die vertraglichen jährlichen Kosten sind bei DOI 50$ weniger als bei handle. Die Kosten die über den Vertrag, bzw. die Registrierung hinaus entstehen, unterscheiden sich in dem Maße, wie die betreibende Institution formelle Qualitätssicherung betreiben will. Die Generierung und Übermittlung vollständiger und korrekter Metadaten bedeutet finanziellen Aufwand für die Institution (bibliothekarisches und technisches Personal). Aus der Qualitätssicherung entstehen jedoch auch Mehrwertdienste für WissenschaftlerInnen und die betreibende Institution, wie z.B. Nutzungsstatistiken und Linked Data.

Aufgrund der bereits bestehenden Mehrwertdienste, der Verbreitung und Qualität wird trotz der durch die Qualitätssicherung entstehenden Kosten das DOI System als zukünftiges PIS für Forschungsdaten an der Humboldt-Universität zu Berlin empfohlen.

Eine mögliche Lösung wäre auch die Nutzung von handle und DOIs, sobald die Nutzungsszenarien von Forschungsdaten geklärt werden und die jeweiligen Kategorien von Forschungsdaten für das dafür geeignete PIS selektiert werden können.[12]

---

[10] S. dazu Beispiellösung beim DataverseNetwork (Uni Harvard): http://thedata.org/citation/standard

[11] Zur Verbreitung von DOIs siehe auch http://www.crossref.org/01company/crossref_indicators.html

[12] Für die Erweiterung der Nutzung von handles mit DOIs im DataverseNetwork (DVN) spricht sich auch Harvard Uni aus: “Since DOIs are based on handles, it is feasible to extend the DVN software to also support DOIs as the persistent identifiers in the data citation. Once the pricing structure issuing DOIs at a variety of scales has evolved, one can easily imagine supporting some collections using handles and some collections using DOIs within a DVN.” (Stand: Jan 2011, http://dx.doi.org/10.1045/january2011-crosas)

Offene Fragen

Diese Empfehlung trifft für Forschungsdaten zu, die an einer Institution wie der Humboldt-Universität **qualitätsgesichert** vorgehalten werden sollen. Für Forschungsdaten, zu denen entweder keine Metadaten existieren oder für die keine Metadaten vorgesehen sind, stellt jedoch handle eine pragmatische Alternative dar.

Eine Institution sollte dahingehend zunächst klären, zu welchem Zweck die Forschungsdaten mittels PI von WissenschaftlerInnen referenziert werden (sollen)? Hierbei ist zu unterscheiden ob die persistente Adressierung von Forschungsdaten lediglich internen Arbeitszwecken dient, die keine qualitätsgesicherten Metadaten voraussetzen, oder ob mit Forschungsdaten qualitätsgesichert (z.B. durch die Verbindung zu einer Publikation) eine globale Referenzierung und Zitierung ermöglicht werden soll. Der erste Fall könnte für eine Institution zur Folge haben, dass enorme Datenmengen vorgehalten werden, die nicht durch Metadaten beschrieben sind und dadurch ihre Nachnutzbarkeit durch heutige oder zukünftige Dritte unmöglich macht. **Es stellt sich somit die zweite zu klärende Frage, ob die an einer Institution vorgehaltenen Forschungsdaten an eine inhaltliche Qualitätssicherung gebunden sein sollen oder nicht?**

Sobald die Anwendungsszenarien von Persistent Identifiern an der Humboldt-Universität mehrheitlich geklärt sind, könnte zusätzlich zur Vergabe von DOI die Vergabe von handles erfolgen. Diese Parallelstrategie, ausgerichtet auf die unterschiedlichen Zwecke, birgt jedoch auch einen erhöhten technischen, personellen und daher finanziellen Aufwand.

Anhang: Tabelle (Stand: Januar 2013)

Aspekte	Handle System	Digital Object Identifier (DOI)
Kosten	50$ jährlich + 50$ einmalig oder 425$ für 10 Jahre + 50$ einmalig	kostenfreie DOI-Registrierung für akademische Einrichtungen mit Hauptgeschäftssitz Deutschland, sonst 150€ jährlich
Verbreitung	200.000 Präfixe, welche Benutzer erlauben handles zu nutzen	60 Millionen
Infrastruktur (eigener Server)	Betrieb eines eigenen lokalen Handle Servers notwendig → Verfügbarkeit muss auf Dauer sichergestellt werden	keine eigene notwendig
Personal-aufwand	Installations- und Administrationsaufwand, für den Service kein Entwicklungsaufwand, da Software frei verfügbar	fortlaufender Kostenfaktor für den CMS bei Nutzung von DOI bei Erschließung der Metadaten für bereits vorhandene Forschungsdaten
Identifizierung / Registrierung	Zentrales Handle-Registry für die Präfixe	- zentrale Registrierung von Diensten, - Nutzer müssen sich bei den Serviceagenturen registrieren und den Vertrag unterschreiben
Links	http://www.handle.net/index.html	http://www.doi.org/factsheets/DOIKeyFacts.html
Qualitäts-sicherung	vertraglich nicht vereinbart; Verantwortung des Datenzentrums	vertraglich zwischen DataCite-Serviceagentur und Datenzentrum vereinbart
Anmerkungen	Empfehlung der CLARIN-D Initiative für das Projekt LAUDATIO	basiert auf handle
Beispiel	<handle> ::= <handle prefix> "/"<handle suffix> z.B.: hdl:1902.1/19160, die URL dazu: http://hdl.handle.net/1902.1/19160	<doi>:10.<doi präfix> "/"<doi suffix> z.B.: doi:10.4232/1.1, die URL dazu: http://dx.doi.org/10.4232/1.1

Anwender-beispiele	DOI-Anwender, DSpace-Anwender, Library of Congress, DataverseNetwork (Unis in US, NL), eSciDoc, Biodiversity Heritage Library	10 Registrierungsagenturen + untergeordnete Serviceagenturen; die größten international agierenden sind CrossRef (für Zeitschriftenartikel und Bücher) und DataCite (für Forschungsdaten). // CrossRef-Nutzer sind Bibliotheken (>970, auch LoC) und Verlage (>1528); DataCite-Nutzer sind Hochschulen und wissenschaftliche Einrichtungen, darunter LMU, ETHZ, KIT (s. "Registrations by Datacenters" http://stats.datacite.org/)
Mögliche Konsequenzen für die HU als Datenzentrum	Einrichtung eines eigenen lokalen handle-Servers notwendig, fortlaufender Administrationsaufwand des handle-Servers; Aufwand für periodische Pflege der Identifier; Aufwand für Anpassung der angestrebten Repositorylösungen (z.B. Forschungsdatenrepository LAUDATIO, Bilddatenbank) zur Entwicklung der Schnittstelle	Metadaten müssen Mindestangaben beinhalten, wodurch ein redaktioneller Aufwand bei jedem einzelnen Datensatz und fortlaufende Kosten für den CMS bzw. der HU entstehen; Aufwand für Anpassung der angestrebten Repositorylösungen zur Entwicklung der Schnittstellen zur Serviceagentur